DOCUMENT RESUME

ED 405 326 SP 037 209

AUTHOR Tanner, David E.
TITLE The Long (Suit) and the Short (Comings) of Authentic Assessment.
PUB DATE 97
NOTE 17p.; Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education (49th, Phoenix, AZ, February 26-March 1, 1997).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Construct Validity; *Content Validity; Elementary Secondary Education; Evaluation Criteria; Evaluation Methods; *Interrater Reliability; Measurement Techniques; *Performance Based Assessment; Self Evaluation (Individuals); *Student Evaluation; Testing
IDENTIFIERS *Authentic Assessment; Performance Indicators

ABSTRACT

During a period when the reform-minded are very critical of the degree to which testing is actually related to the conditions for which data are employed, authentic assessment offers the opportunity to evaluate learning in settings closely related to the real world. It also allows the evaluator to tailor assessment conditions for individual learners and circumstances. However, it also poses a number of problems. Judgment reliability, or what was conventionally termed inter-rater reliability, may be a problem and, to the degree that assessment tasks are tailored to individual learners, content validity may also be problematic. To be authentic, the assessment system must include criterion-based standards, provide multiple indicators of performance quality, and reflect judgment reliability. (Contains 11 references.) (Author/ND)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 405 326

The Long (suit) and the Short (comings)

of Authentic Assessment

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
IN OTHER THAN PAPER COPY
HAS BEEN GRANTED BY

D Tanner

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

AACTE, Phoenix, 1997

David E. Tanner

California State University Fresno

SO 027209

BEST COPY AVAILABLE

Abstract

During a period when the reform-minded are very critical of the degree to which testing is actually related to the conditions for which data are employed, authentic assessment offers the opportunity to evaluate learning in settings closely related to the real world. It also allows the evaluator to tailor assessment conditions for individual learners and circumstances. However, it also poses a number of problems. Judgment reliability, or what was conventionally termed inter-rater reliability, may be a problem and to the degree that assessment tasks are tailored to individual learners, content validity may also be problematic.

A close examination of the tests which classroom teachers administer suggests that they sometimes offer poor content coverage, emphasize a narrow range of cognitive skills, and can be substantially removed from experiences that the students need to prepare for (Kamphaus, 1991). Many of those disenchanted with traditional testing as the primary student assessment mechanism, have turned to some more authentic type of assessment as an alternative. Noted Darling-Hammond, Ancess, and Falk (1995),

> Increasingly, local schools, school districts, and states are experimenting with . . . alternatives to standardized testing for assessing student learning and performance. Persuaded that traditional standardized tests fail to measure many of the important aspects of learning, and that they do not support many of the most useful strategies for teaching, practitioners are introducing alternative approaches to assessment into the classroom--approaches that help teachers look more carefully and closely at students, their learning, and their work. (p. 4)

Where traditional assessment may involve the use of instruments which are unique to the classroom setting, "authentic assessment" refers to evaluating learners' competency with tasks that are worthwhile and meaningful in the classroom and beyond.

The concept isn't new. Physical education, art, music, and many vocational education classes have a tradition of using authentic work samples to assess student performance. The approach is much less common, though certainly not unknown, in traditional academic subjects and it has come to the fore as a part of the education reform movement. Authentic assessment in the academic areas, noted Elliott (1991), " is a central piece of the outcome oriented education and school restructuring movements in the United States" (p. 274).

Advocacy in education is often disconnected from critical and analysis. However enticing authentic assessment may be, emerging research suggests that there are questions relating to the quality of authentic assessment which

have not yet been answered. This paper is an analysis of authentic assessment's strengths, and also some of its potential weaknesses.

Authenticity

By their nature, most traditional tests contain an element of artificiality compared to the setting in which the learning will ultimately be used. Fitzpatrick and Morrison (1971) called the real-world setting, the **criterion situation**. Authentic assessment tries to emulate the criterion situation for the assessment activities and produce "a genuine rather than a contrived learning experience that provides both the teacher and student with opportunities to learn what the students can do" (Darling-Hammond, et al., 1995, 3-4).

Much of the appeal that authentic assessment has is based in the consistency between the assessment, and the activity the task for which the learner is being prepared. A multiplication problem posed so that the learner must determine how much interest would accrue in a year on a $9,000 loan at 12% is probably more authentic than a problem presented simply as the solution to .12 X 9,000. However, it is less authentic than the task would be if someone were sitting at a loan officer's desk trying to determine whether she could afford to borrow $9,000 at 12%. Rather than placing assessment in an authentic/not authentic dichotomy, authenticity is better viewed in relative terms.

Participants in a teacher preparation course might be asked to complete any of the tasks in Figure 1 as they learn to prepare instructional units. Each task is logically related to preparing instructional units, and each uses a slightly different approach and provides for a different level of authenticity. The various tasks occupy positions along what might be termed a "fidelity continuum" from those that are precisely the same as others learners will be expected to complete after leaving the classroom, to others so unrelated to real world

situations as to be completely contrived.

Figure 1

Educational Tasks with Varying Degrees of Authenticity

1) Complete a multiple choice quiz with items based on what is good and what is poor practice in preparing instructional units.

2) Write an essay on the "dos" and "don'ts" of preparing units.

3) Outline an instructional unit which meets appropriate criteria.

4) Complete a critique of an instructional unit.

5) Prepare an instructional unit.

6) Prepare and field-test an instructional unit.

It's important not to make assumptions about authenticity based only on the structure of the evaluation. Some portfolio entries, for example, are no more authentic than the paper-and-pencil assessments they replace. Preparing an instructional unit for a sixth grade class is certainly an important task and in certain instances it's an authentic task, but if it is intended to reflect how well someone can teach a unit to a sixth grade class, it lacks authenticity. Furthermore, sometimes paper-and-pencil tests are *not* artificial since many of the things we have to do, such as completing income tax returns, legitimately involve paper-and-pencil activities.

Principles of Authentic Assessment

Authentic assessment is part of a system which includes more than an authentic setting. Newmann and Archbald (1992) identified three 'principles of evaluation' in an authentic assessment program. In order to be effective, the

assessment system must include **criterion-based standards**, provide **multiple indicators of performance quality**, and reflect **judgment reliability**.

Criterion-Based Standards

The need for criterion-based standards is based in the requirement that those conducting the assessment define what the minimum levels of acceptable performance are. Since there are usually different levels of acceptable learner performance, assessors must also develop a mechanism for explaining variations in performance quality.

The need for the set of evaluative criteria suggests a link between authentic assessment and the criterion-referenced assessment of the 1960s. The earlier movement had its origins in behavioral psychology which lauded the safety inherent in learner activities which could be observed directly and which required very few inferential conclusions . While most advocates of authentic assessment do not recognize a relationship to criterion-referenced testing, the use of an objective set of performance criteria is common to both movements.

Figure 2 provides the standards for a seventh grade science project which learners are to develop after studying the earth's geology for several weeks. Students are encouraged to produce a variety of projects which each allow them to demonstrate what they have learned, but each must meet specific standards reflected in the scoring criteria.

Because criteria are established is no guarantee that the criteria are valid. One of the lessons from the criterion-referenced testing of the earlier reform period is that the need to establish specific criteria can lend itself to trivializing the curriculum because those concepts which are easiest to objectify become the most natural targets for assessment. They are easier to assess than are the learners' grasp of abstractions. One could ask any number of questions

regarding order in which the planets occur in the solar system, for example, but whether the topic deserves such attention could be a quite different question.

Figure 2
Criteria for Assessing a Science Project

Scoring Protocol:
If the best description of the element is,
"very good," the score for that element should be.....18-20
"good," the score should be.....15-17
"fair," the score should be.....12-14
"unacceptable," the score should be.....<12

A. Degree to which the project reflects important geologic concepts ___

B. Creativity ___

C. Clarity ___

D. Visual appeal of the project ___

E. Effort ___

Total ___

Interpretation:
The project is acceptable if the score for A is equal to or greater than 14 and the total score is at least 70 points, and there are no individual scores lower than 12. Should any of these conditions not prevail, the project is unacceptable and must be resubmitted.

Multiple Indicators of Quality

The need for multiple quality indicators recognizes that nearly any academic task offers more than one skill or ability to be assessed. Although we traditionally sum up a learner's performance in a letter grade, or a brief verbal description ("pass," or "credit") those are summary assessments which represent

a number of discrete skills and abilities. They are aggregated grades which often communicate very little about the individual component elements of the learner's performance. A unit of instruction from American History would certainly involve recall ability. It also probably involves a substantial amount of listening, reading, and writing abilities. It could conceivably require research and critical analysis, a grasp of basic economics principles, and a synthesis of previously (for the learner) unrelated bits of information. If the learner receives a C on the unit and we accept that the operational definition of C is "average," what does that reveal of the learner's performance in the component abilities, and what does 'average writing ability,' to take one example, mean?

Single indicators aren't always misleading or inappropriate. Applicants need to know whether or not they are admitted to the university they applied to, an outcome which can be addressed by a single indicator, but authentic assessment makes an issue of recognizing that ordinary learning activities usually call for something more communicative that a single, summary response.

In addition to the multiple quality indicators associated with a complex product, there also need to be multiple products to assess. Newmann and Archbald (1992) note,

> In making overall assessments about individuals' accomplishments, it is usually necessary to consider not just one performance (even if it is judged by multiple criteria), but instead to characterize a variety of performances over time. Just as it becomes more informative to examine a full portfolio of work, an athlete's record over a whole season, or several tunes in a musician's repertoire, the assessment of students will be more authentic if it permits observation of patterns of strengths and weakness over a sustained period. (p. 81)

Judgment Reliability in Authentic Assessment

In traditional settings, test reliability refers to the degree to which the assessment data are consistent when the conditions under which the test is

administered (the environmental setting, the range of ability levels in separate groups of test-takers) are consistent. Aside from characteristics related to the respondents and the setting, an important potential source of reliability difficulty is variability in the way the assessor responds to assessment data. In simplest terms, in order to have judgment reliability, a response which is judged adequate, or correct for one test-taker must be judged equally adequate or correct for any other learner when the quality of the response is the same. Factors which are irrelevant to the quality of the response must not influence the judgment.

Traditional testing minimizes reliability problems by limiting the human element in the assessment. This is fostered by developing tests items which can be discretely scored either right or wrong. There are usually not degrees of adequacy or correctness, the respondent either does or does not score the item correctly. These conditions minimize subjectivity and maximize the potential for achieving consistent (reliable) scoring results.

In an authentic assessment system a number of conditions prevail that make environmental conditions variable. Because there is often no common test as such, neither is there a common setting for assessment. Authentic assessment calls for multiple evaluative criteria and the materials that learners produce to meet those criteria may be substantially different from one learner to the next; often they are encouraged to be. Fitzpatrick and Morrison (1971) claim that to the degree that circumstances are controlled so that different learners' situations are comparable, elements of the fidelity of the setting--its 'authenticness' are probably forfeit.

Part of the advantage of multiple criteria is that learners have flexibility. They are to be able to satisfy the standards with a variety of projects and activities. The absence of a common setting and even, perhaps, of common

materials requires that reliability be approached in terms of the degree to which several raters agree on the same performance, a process which places a premium on the judgment skills of the assessors (Mehrens, 1992).

Traditionally, multiple measures of learners' projects have traditionally suffered from low inter-rater reliability and achieving even moderate levels can be difficult and relatively expensive, but it can be done. When the assessment is directed at a multiple component product such as a learner's portfolio, Reckase (1995) maintains that inter-rater reliability is highest when the individual components of the portfolio are scored and then a composite score is created from the individual elements. Using this approach, he achieved a theoretical inter-rater reliability high enough to be useful.

An alternative to Reckase's (1995) approach is **holistic scoring** which is scoring the project only as a complete entity. Another is to score the individual components of the project without aggregating the individual scores. Both procedures tend to reflect a great deal of scoring variability.

In addition to contributing to scoring reliability, the procedure of scoring individual components and then aggregating scores serves another purpose. Individual component scores allow detailed assessment of the learner's performances on the individual project components while still yielding a summary judgment.

Validity in Authentic Assessment

Issues of reliability and validity generally are more associated with traditional standardized testing than with authentic assessment, but there are issues which transcend the type of assessment employed.

Assessment Relevance, and Consequential Validity

Advocates of authentic assessment rarely raise the content, criterion-related, and construct validity issues so germane to traditional assessment.

Rather the discussion is of **assessment relevance**, the degree to which the assessment is related to what the learner is being prepared to do beyond the particular assessment setting, and the assessment's **consequential validity**, "the consequences for students and for schools of using a particular form of assessment" (Darling Hammond, et al., 1995, p. 64). By these definitions, if an assessment fits with what learners are being prepared to do beyond the classroom, and if it strengthens teaching and learning, it is valid. What often occurs is that assessment relevance addresses many of the issues common to predictive validity and, to the degree that consequential validity is related to a specific intellectual skill that the learners develop, it may also address construct validity.

Consider a situation in which a learner develops an exhibit of labeled rocks as the product of a physical science unit dealing with the geology of the region. A collection of rocks certainly appears to be a more authentic product than a score from paper-and-pencil test of how well the learner grasps the characteristics of igneous, sedimentary, and metamorphic rocks. The validity question, however, is whether the assessment of the exhibit is more relevant than the test score *to the learner's grasp of the area's geology* and whether the kind of learning required to complete the project are valued outcomes. They aren't easy questions to answer. The consequential validity question can only be answered by comparing the kind of teaching and learning that resulted in the rock collection to that which would have occurred in the more conventional setting. Simply setting aside procedures which rely heavily on conventional testing is no guarantee that a non-testing substitute provides greater validity in spite of what appears to be a more authentic product and a more authentic setting.

Employing Common vs. Individual Standards

When learning and assessment activities are planned, one anticipates what learners must accomplish and then designs accordingly. Ideally, there is harmony between the behavior the assessment calls for and the behavior the learner exhibits. On this general level, authentic assessment and traditional testing are in essential agreement. A question which arises is whether what is to be accomplished is the same for all learners. Berlak (1992) argued against a common standard noting that one of the problems with traditional testing is the assumption that what is identified as the relevant learner behavior can be generalized to all learners.

> . . . the assumption that there can be a meaningful nationwide statewide, district-wide, or even school wide consensus on the goals of schooling and on what students should learn and how they should learn is untenable. . . In a multicultural society which values difference, consensus is undesirable. (p. 14)

So while probably no one argues against the need to determine what must be accomplished as a result of the learning exercise, it is in Berlak's logic, potentially a different accomplishment for each state, district, school, and perhaps by extension, for each learner.

This is an important point of departure for advocates of authentic assessment. Traditional assessment techniques sustain an objective set of quite specific goals and objectives which are intended for all who fall within a particular age or grade classification. Authentic assessment advocates the need for standards that are responsive to situational and individual differences, and that are general enough that a variety of activities can be mechanisms for meeting them. This is particularly important since learners frequently are required to construct original responses (Mehrens, 1992).

Conclusions

Sometimes movements are defined by what they aren't, which may be the case with authentic assessment. Archbald (1991) noted that it, "appears to describe just about any alternative to a standardized multiple-choice test" (p. 279-280). When the increased testing which characterized the early part of the current reform period did not sufficiently alter performance in the schools, standardized testing appeared not to be part of the answer, and the reform movement expanded to reflect a distrust of tests. They were an easy target. Kamphaus (1991) reported test's sampling of content domains in some instances to be "nothing short of appalling" (p. 301), but content coverage wasn't the only issue. Traditional school assessment procedures often did not address problem-solving ability and other intellectual skills substantially beyond the level of recall.

Having noted some of the criticisms that have prompted the rise of authentic assessment, we ought not to assume that the positions taken by advocates necessarily reflect a substantial research base. Authentic assessment is a movement in process and both its role and its relationship to standardized testing are still being defined. Archbald's (1991) position is that standardized testing's role is primarily in the area of individual student diagnosis. Interestingly, that is the role that Mehrens (1992) defined for authentic assessment.

Although some of the language has changed, many of the concepts associated with authentic assessment are not new. Elliott (1991) recognized that the methods employed in authentic assessment are clearly related to what was called "applied behavior analysis" 30 years ago. That movement spawned criterion-referenced testing which in many ways is a precursor to authentic assessment in spite of the fact that very little effort has been made to cross reference current procedures with the earlier work. Beyond that theoretical link,

the idea of placing assessment in as authentic a setting as possible has long been a fixture in some disciplines. The effect of the current movement is to expand the concept into disciplines where it has not historically been housed.

Authentic assessment reflects more than an approach to student assessment. It presumes that students will produce something that reflects not a narrow, compartmentalized repetition of what was presented to them, but an integrated scholarship which connects their learning to learning housed in other disciplines and which is presented in a setting consistent with that in which the learning is likely to be most useful in the future. Those objectives are difficult to argue with and indeed, reflect outcomes which are highly regarded by all educators.

Mehrens (1992) suggested that part of authentic assessment's strength is in the greater ease with which it can be integrated into instruction and become part of the larger educational picture. Perhaps any kind of assessment drives the associated instruction to some degree, but authentic assessment's focus on a range of skills and abilities probably provides a sound basis for instructional decision-making. This is a theme that Christenson (1991) took up in her critique of authentic assessment when she noted,

> Authentic assessment models good instruction. The concept of authentic assessment is reified by thinking of our knowledge base on instructional effectiveness. By determining how to assess students' performance on meaningful tasks, the probability that instruction will be modified is high. In this way, authentic assessment may be the stimulus for implementing the instructional effectiveness guidelines more systematically. (p. 297)

Authentic assessment has much to offer but it does not come without drawbacks. Christenson (1991) notes that it is time and cost intensive. It requires that educators identify the essential learning competencies by grade level and discipline, and in some cases across disciplines so that the synthesis

that advocates prize can occur. One of its great strengths is the attention paid to content validity, but other elements of validity are not given similar attention in authentic assessment. Kamphaus asked, "What aspect of psychometric practice is addressed by authentic assessment? Is this approach aimed at increasing the predictive, concurrent, or factorial validity of educational achievement measures? I do not get a clear sense that this is the case" (p. 300-301).

Because criterion-related validity and factorial validity have been associated with traditional testing doesn't obviate the need to establish their currency in authentic assessment, but authentic assessment supporters have done little to address these elements. Instead authentic assessment advocates appeal to the authenticness of the setting as the presumption of validity but this is probably inadequate. There is still a need to demonstrate that performing what is described as an authentic task in the school setting indicates how well one will perform when the learning is called upon later.

There also needs to be some assurance, besides calling the assessment 'authentic,' that the assessment of the learner's performance on the task reflects the degree to which the skills and abilities involved in the task have actually been mastered. Perhaps an appropriate closing comment is offered by Kamphaus who noted, "authentic assessment is aimed at improving content validity. It's not a technological breakthrough, it's a complementary . . . tradition aimed, with others, at improving educational assessment practice" (p. 302).

References

Archbald, D.A. (1991). Authentic assessment: Principles, practices, and issues. *School Psychology Quarterly, 6*, pp. 279-293.

Archbald, D.A. and Newman, F.M. (1992). "Approaches to assessing academic achievement," in Berlak, H., Newmann, F.M., Adams, E., Archbald, D.A., Burgess, T., Raven, J., and Romberg, T.A., *Toward a new science of educational testing and assessment* (pp. 139-180). New York: State University of New York Press.

Berlak, H. (1992). "The need for a new science of assessment," in Berlak, H., Newmann, F.M., Adams, E., Archbald, D.A., Burgess, T., Raven, J., and Romberg, T.A., *Toward a new science of educational testing and assessment* (pp. 1-21). New York: State University of New York Press.

Christenson, S.L. (1991). Authentic assessment: Straw man or prescription for progress? *School Psychology Quarterly, 6*, 294-299.

Darling-Hammond, L., Ancess, J, and Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. National center for restructuring education, schools, and teaching, Teachers College, Columbia University: New York.

Elliott, S.N. (1991). Authentic assessment: An introduction to a neobehavioral approach to classroom assessment. *School Psychology Quarterly, 6*, pp. 273-278.

Fitzpatrick, R. and Morrison, E.J. (1971). Performance and product evaluation in E.L. Thorndike (Ed.). *Educational Measurement* (2nd Ed.) pp. 237-270. Washington D.C.: American Council on Education.

Kamphaus, R.W. (1991). Authentic assessment and content validity. *School Psychology Quarterly, 6*, pp. 300-304.

Mehrens, W.A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11* (1), pp. 3-9, 20.

Newman, F.M. and Archbald, D.A. (1992). "The nature of authentic academic achievement," in Berlak, H., Newmann, F.M., Adams, E., Archbald, D.A., Burgess, T., Raven, J., and Romberg, T.A., *Toward a new science of educational testing and assessment* (pp. 71-83). New York: State University of New York Press.

Reckase, M.D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14* (1), pp. 12-14, 31.